DOCUMENT RESUME

ED 441 870 TM 031 212

AUTHOR Tanguma, Jesus
TITLE Steps in Test Construction.
PUB DATE 2000-04-00
NOTE 23p.; Paper presented at the Annual Meeting of the Southwestern Psychological Association (Dallas, TX, April 20-22, 2000).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Test Construction; Test Content; Test Format; Test Items
IDENTIFIERS Test Specifications

ABSTRACT

This paper addresses four steps in test construction specification: (1) the purpose of the test; (2) the content of the test; (3) the format of the test; and (4) the pool of items. If followed, such steps not only will assist the test constructor but will also enhance the students' learning. Within the "Content of the Test" section, two examples of tables of specifications are presented. In addition, detailed guidelines are provided for writing different item formats. (Contains 2 tables and 10 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 870

Running head: STEPS IN TEST CONSTRUCTION

Steps in Test Construction

Jesus Tanguma

Texas A & M University

J. Tanguma

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM031212

Paper presented at the annual meeting of the Southwestern Psychological Association, Dallas, April 2000.

BEST COPY AVAILABLE

Abstract

This paper addresses four steps in test construction specifying a) the purpose of the test; b) the content of the test; c) the format of the test; and d) the pool of items. If followed, such steps will not only assist the test constructor, but also enhance the students' learning. Within the content of the test section, two examples of table of specifications are presented. Also, detailed guidelines for writing different item formats are provided.

Steps in Test Construction

Tests and the use of test results have been with us since the beginning of recorded history. The Bible (Judges 12: 5-6) shows how the Gileaditis orally questioned the Ephraimites concerning their nationality:

> And the Gileaditis took the fords of the Jordan against the Ephraimites. And when any of the fugitives from Ephraim said, "Let me go over," the men of Gilead said to him, "Are you an Ephraimite?" When he said "No," they said to him, "Then say Shibboleth," and when he said "sibboleth," for he could not pronounce it right; Then they seized him and slew him at the fords of the Jordan. And there fell at that time forty-two thousand of the Ephraimites. (p. 259)

Most students, if not all, have recently been administered a test of some kind. As Sax (1974) noted, "a test may be defined as a task or series of tasks used to obtain systematic observations presumed to be representative of educational or psychological traits of attributes" (p. 3). However, although test results are used in a wide variety of ways, many test writers, at the public school level, are not well versed in the construction of exams. Consequently, the test items they write may not measure what the items were intended to measure. Thus, the scores obtained from such test items may neither be valid nor reliable. The purpose of this paper is to present steps in the test construction process which, if followed, will not only assist the test constructor, but will also enhance student learning.

Once a decision has been made to write a test, it is necessary to plan the test so that it will provide the most useful information. To achieve this, a blueprint (plan) of the test may be designed. Such blueprint should include, but not be limited to, the purpose of the test, a table of

specifications, the type of test to be constructed, a pool of items to select from, and an item analysis. This paper addresses the purpose of the test, presents two examples of tables of specifications, describes and provides detailed guidelines for writing different item formats, and addresses the building of a pool of items.

Purpose of the Test

Describing what specific construct the test will measure, how the results will be used, and who will take the test provides focus during the test construction process and a framework for evaluating the completed instrument. In deciding what the test will measure, the test developer needs to know what output the students are to produce. For example, if a teacher wants to know whether the students know how to factor, then the test developer will write out problems asking students to factor. Likewise, if a high school principal wants to know whether to place the incoming freshmen in an Algebra I or a remedial math course, he would ask the test developer to design a placement test. "Summative testing of course objectives can generally be best implemented by teacher-made surveys of course content, product evaluations, or performance tests" (Thorndike, Thorndike, Cunningham, & Hagen, 1991, p. 194).

Another important item that must be taken into consideration when writing a test is how the results of testing are to be used. Test results may be used to decide who is admitted into an educational institution. For example, the GRE scores are part of the general criteria for admission to Texas A & M University. Once the individual is admitted, a placement test might be administered to such subject. The student's advisor might use the results of such a test to place the student in the most appropriate classes. Other tests are used to decide who is hired among a group of applicants or to decide who is to be promoted. Tests are also used to evaluate the

effectiveness of a program. These tests are administered at the beginning as well as at the end of the program of study. In such a case, the scores themselves are not as important as the difference between the pretest and the posttest (Wiersma & Jurs, 1990, p. 28).

In making administrative decisions such as selection, classification, and curriculum planning, educational tests are used. An example of this might be that of selecting a TAAS, Texas Assessment of Academic Skills, program to be used by a school district in trying to improve their TAAS scores. Here, the administrators are to select the program that is most practical for the district.

Educational psychologists develop and administer tests in their search of new knowledge. That is, educational psychologists test their own research in an attempt to create new knowledge. As Sax noted in 1974 .

> Some tests specialists construct new and useful educational and psychological tests; others develop mathematical formulas that lead to better understanding of assumptions and interrelationships among different factors and variables. Still others seek new theories of personality, attitudes, or intelligence by studying how different groups of people respond to tests of different types. Thus in addition to having a practical bent, the study of measurement creates its own body of knowledge. (p. 14)

In the public schools, "the major purpose for using teacher-made test in the classroom is, and should always be, to improve instruction" (Bott, 1996, p. 27). When used in the classroom setting, test results can be used to determine if the students have learned the subject matter. Then, the teacher can make a decision whether to advance to the next unit of instruction or to re-teach. If re-teaching is needed, a careful analysis of the work turned in by the students might

indicate the specific areas of need as well as the methods that will be most appropriate for each student. Test results can also be used for grading. "The primary function of grading and marking is to communicate effectively to a variety of audiences the degree of achievement of academic competence of individual students" (Oosterhof, 1994, p. 334).

Content of the Test

Tables of specification should be prepared before the beginning of instruction. Therefore, once the purpose of the test has been identified, the next step is to develop a table of specifications for the test.

> A table of specifications ensures that the test has a proper balance of emphases; this will guide the test developer as blueprints and specifications guide the building contractor. It is valuable to indicate not only the various objectives in mind but also, at least roughly, the relative amount of emphasis on each objective. (Hopkins & Stanley, 1981, pp. 176-177)

Tables of specification vary as to how detailed they are. "A good rule of thumb to follow in determining how detailed the content area should be is to have a sufficient number of subdivisions to ensure adequate and detailed coverage. The more detailed the blueprint, the easier it is to get ideas for test items" (Mehrens & Lehman, 1984, p. 68). "The numbers within the table of specifications indicate the number of test questions to be associated with each content area and capability. Larger numbers indicate more emphasis being given to a particular content area and capability" (Oosterhof, 1994, p. 55). Notice that Table 1 only indicates the number of test items per objective. It neither specifies the weight of each item nor that of the entire objective. Moreover, Table 1 does not relate the test questions to Bloom's (Kubiszyn &

Borich, 1984, p. 52) taxonomy. The column labeled "Total," in Table 2, indicates the total number of questions from a specific objective as well as the percentage of the test devoted to such objective. Additionally, the row labeled "Totals," also in Table 2, indicates the total number of questions from the test as well as the percentage of the test that relate to Bloom's (Kubiszyn & Borich, 1984, p. 52) behavioral objectives. Table 1 shows a very general table of specifications on a unit on quadratic equations. Table 2 shows a more detailed table of specifications. Such a table will not only guide the test developer in a more systematic construction of the test, but will also inform prospective examinees how to best prepare for the test. In doing so, the teachers are forced to write a more balanced test and students will not be able to make comments such as "the material we were tested on wasn't covered in class" (Mehrens & Lehman, 1984, p. 67). In a way then, "the specs can help provide for optimal learning on the part of the pupils and optimal teaching efficiency on the part of the instructor" (Mehrens & Lehman, 1984, p. 68).

Because the terms reliability and validity will be used throughout this paper, they need to be defined. "Reliability of measurement is consistency-consistency in measuring whatever the instrument is measuring" (Wiersma & Jurs, 1990, p. 155). "Validity pertains to the degree to which a test measures what it is supposed to measure" (Oosterhof, 1994, p. 53).

Format of the Test

Once the purpose of the test as well as what the test is to measure have been decided, the next step is the preparation of the test format. "There are basically two kinds of item formats: objective (matching, true-false, and multiple-choice) and essay (sometimes called created-response and which consists of completion items and the brief and extended essay)" (Mehrens & Lehmann, 1984, p. 74). Essay type items require the student to select, arrange, organize, and

express ideas or to produce original solutions to problems. Thus, "essay items should not ask for definitions or list of information, which require only recall behavior of the students" (Wiersma & Jurs, 1990, p. 71). In deciding which item format to use, one of the factors the test developer needs to consider is the age of the persons to be tested. While most high school students might appreciate a well-written multiple-choice test, pre-k students will not benefit from such a test format. Another factor to consider is time. For example, while a multiple-choice test is graded very quickly, the writing of such questions could be very time consuming. In contrast, an essay question might be very easy to write, but will require a lot of time to grade. Other factors to consider include the number of people to be tested, the testing place, and the teacher's ability to write different types of items (Mehrens & Lehmann, 1984, p. 76). Consequently, the test developer needs to be careful in selecting the item format(s) to be used in a given test. A description of different item formats follows.

True-false items are essentially statements to which the student responds true or false, yes or no, or right or wrong. "The yes-no format is often used to measure attitudes, values, beliefs, and interests. The right-wrong or yes-no varieties are more useful for testing young children, who are better able to comprehend the concept of right-wrong than true-false" (Mehrens & Lehmann, 1984, p. 141). Such items are very easy to construct, to grade, and, "because true-false items are short, they can sample a considerable amount of content within a single test" (Oosterhof, 1994, p. 153). However, true-false items encourage guessing because there is an equal chance that either answer will be correct, and more likelihood of a correct guess than on a multiple-choice item. That is, true-false items may be answered correctly without any knowledge of the subject matter being tested because of grammatical clues. Thus, producing scores whose

reliability is very low. "In fact, on an item-per-item basis, true-false tests tend to have the lowest reliability" (Mehrens & Lehmann, 1984, p. 145).

The following are some guidelines from Thorndike et al. (1991) in writing true-false items. First, make sure that the item is unequivocally true or false. Second, avoid the use of specific determiners (an unintentional clue to the correct answer). Third, avoid the use of negative statements and particularly double negatives. Fourth, limit each item to a single idea. Fifth, make true and false statements approximately equal in length. Sixth, use an approximate equal number of true and false statements.

The following is an example of typical well-written true-false questions.

Directions: Listed below are a number of statements. Some are true and some are false. If the statement is true, draw a circle around the "T" at the left of the statement. If the statement is false, draw a circle around the "F." The first item is answered as an example.

T F x. The systolic pressure is the numerator of a blood pressure reading.

T F 1. The brachial artery is the artery used to palpate for a blood pressure.

T F 2. The adult male's blood pressure may be slightly higher than the adult female's.

T F 3. The waiting period between taking blood pressure on the same patient and the same arm is 2 minutes. (Bott, 1996, p. 62)

The following item illustrates a variation of the typical true-false question. This question requires a higher level of understanding in order to correctly answer the question.

Directions: If the following items are true, draw a circle around the "T" and do no more. If the item is false, draw a circle around the "F" and explain in the blank why it is false.

T F 1. Intelligence is the sum of an individual's many different abilities to learn. In addition, it represents potential capacities.

Explanation:_____

_____ (Bott, 1996, p. 71)

T F 2. Intelligence is believed to be determined largely by environmental factors.

Explanation: _____

_____ (Bott, 1996, p. 71)

Matching items are a variation of multiple-choice items in which the student associates an item in one column with a choice in the second column. However, "the matching format does not require the construction of plausible distractors, which is an advantage over multiple-choice testing" (Nunnally, 1972, p. 55). The student may associate names of individuals with their accomplishments, events with dates, or countries with their capitals. Items are usually listed in the first column and choices in the second column. This type of item is useful "in testing the knowledge of terms, definitions, data, events, and other matters involving simple relationships" (Mehrens & Lehmann, 1984, p. 138). Additionally, matching items can be constructed relatively easily, quickly, and can cover a considerable amount of material in a single test.

The following are guidelines by Sax (1974) in writing matching test items. First, have more options than items. Second, arrange options and items alphabetically or numerically. Third, limit the number of items within each set. Fourth, place the shorter responses in column B. Fifth, provide complete directions. Sixth, place options on the same page. Seventh, use homogeneous options and items.

The following is a sample matching question.

Directions: Column A describes events associated with United States presidents. Indicate which

name in Column B matches each event by placing the appropriate letter to the left of the number in Column A. Each name may be used only once.

Column A	Column B
_____1. Only president not elected to office.	a) Woodrow Wilson
_____2. Delivered the Emancipation Proclamation.	b) Thomas Jefferson
_____3. Only president to resign from office.	c) Abraham Lincoln
_____4. Only president elected for more than two terms	d) Richard Nixon
_____5. Our first president.	e) Franklin Roosevelt
	f) Theodore Roosevelt
	g) George Washington
	h) Gerald Ford

(Kubiszyn & Borich, 1984, p. 70)

The following is a poorly written matching question.

Directions: Match the terms in Column B with those in Column A. Write the matching letter on the space provided.

Column A	Column B
_____ fungus	a) thermometer
_____simple	b) tuberculosis

machine

_____measures of air — c) chlorophyll

pressure — d) athlete's foot

_____prism — e) heat conduction

_____simplest type — f) refraction

of matter — g) lever

_____photosynthesis — h) legume

_____tides — i) barometer

j) element

k) moon

l) erosion

(Nunnally, 1972, p. 168)

Multiple-choice items each consist of a stem, which asks or implies a direct question, and a series of options or alternatives. "All incorrect or less appropriate alternatives are called distracters or foils, and the student's task is to select the correct or best alternative from all the options" (Sax, 1974, p. 88). As the number of options increases, the chances of guessing the correct response decreases, and the reliability of the test items increases.

Multiple-choice items have several advantages over other item formats. Such items are easily and objectively scored. They can also cover a considerable amount of material in one test. Additionally, "students often find multiple-choice questions less ambiguous than completion or

true-false items. Instructors also find it easier to defend correct answers" (Ebel & Frisbie, 1986, p. 160).

While multiple-choice items are very easy to score, it is very difficult to construct such items. Not always can teachers come up with distractors that are plausible but not correct. Another disadvantage of the multiple-choice items is that such items require "the most time for the student to respond, especially when very fine discrimination has to be made" (Mehrens & Lehmann, 1984, p. 155).

The following are some poorly constructed multiple-choice type questions and their corresponding corrections.

1. Poor: Magellan's primary contribution to world culture is that he was the first person to

    a. circumnavigate the globe.

    b. discover the Atlantic Ocean.

    c. land on American soil.

    d. look for the Fountain of Youth.

    Better: Magellan was the first person to

    a. go around the world.

    b. discover the Atlantic Ocean.

    c. land on American soil.

    d. look for the Fountain of Youth (Sax, 1974, p. 91)

2. Poor: The function of the platelets in the blood is to help in

    A. carrying oxygen to the cells.

B. carrying food to the cells.

C. clotting of the blood.

D. fighting disease (Thorndike et al., 1991, p. 230)

Better: Which of the following structures in the blood helps in forming blood clots?

A. red blood cells.

B. lymphocytes.

C. platelets.

D. Monocytes (Thorndike et al., 1991, p. 231)

The following are guidelines for writing multiple-choice test questions as presented by Nunnally (1972). First, the problem should clearly point to the theme of the correct alternative answer. Second, incorrect alternatives should be plausibly related to the problem. Third, correct alternatives should not be consistently different in appearance from incorrect alternatives. Fourth, alternatives should be randomly ordered for each item. Fifth, avoid irrelevant sources of difficulty in the statement of the problem or in the alternatives. Sixth, avoid including material in the problem that is unrelated to the theme of the intended response. Seventh, do not employ alternatives which say "none of the above," "all of the above," "both a and c above," etc. Eighth, avoid grammatical cues and sentence structures that give away the correct alternative. Ninth, use negatives sparingly in problem statements. Tenth, each item should be independent of every other item. Eleventh, ensure that item content relates to important aspects of the subject matter.

Completion or short-answer items consist of a statement with one or more key words missing and blanks left in their place. "The blanks can either be at the end of the item, which makes it a completion item, or be embedded in the statement" (Thorndike et al., 1991, p. 53). Students are required to fill-in the blanks to complete the statement correctly. If the question is one that can be answered very briefly, it may be put in the form of a completed question, followed by a blank space.

Short-answer items test the students' ability to recall information, rather than to recognize it in context. Such items are good to use when students must be able to remember facts, words, or symbols. It is very difficult to guess the correct answer with short-answer items. Consequently, "short-answer tests tend to be more reliable than multiple-choice or true-false test containing the same number of items" (Oosterhof, 1994, p. 98). It is, however, very difficult to create statements that call for only one correct answer. Since short-answer items require the students to supply their own answer, the guessing factor is minimized and the students either know the answers or they don't.

The following are guidelines for writing completion/short-answer test questions as presented by Nunnally (1972). First, use only one or two blanks. An item calling for more than two blanks may, of course, be used legitimately if it is constructed with caution. An example of the legitimate use of multiple blanks is:

The three elements which are essential for combustion are _____ ,_____ , and _____ .

Second, make sure that only one term will sensibly complete the statement or answer the question. Third, leave only important terms blank. Fourth, place the blank space near the end of the sentence. Fifth, avoid repeating textbook phrasing word for word. Sixth, avoid grammatical cues to the correct answer.

The following are some poorly constructed short-answer questions and their corresponding corrections.

1. Poor: If two angles sum to 120 degrees, the triangle is called an _____ triangle.

   Better: If two angles sum to 120 degrees, the triangle is [a, an] _____ triangle.

2. Poor: Essay items can evaluate a student's ability to _____ ideas in writing.

   Better: Which item format can measure student's ability to communicate ideas in writing? _____ (Oosterhof, 1994, pp. 105-106)

As Mehrens and Lehmann (1984) pointed out, "with the exemption of the oral test, the essay is the oldest test format in use today." An essay item is one for which the student makes a comparison, writes a description, or explains certain points on which instruction has been given. Thus, essay items assess the student's ability to communicate ideas in writing.

Essay items have several advantages over test questions written in other formats. For example, "because essays involve recall, there are no options to select from, and guessing is eliminated" (Sax, 1974, p. 117). Another advantage of the essay items is that such items are relatively easy to construct. However, essay items are time consuming to score.

Since students, in answering essay questions, are to supply their own answer their own way, responses to essay items may be subject to bluffing. Consequently, scores obtained from essay items have low reliability. Moreover, essay items are time-consuming to score, test only a small portion of the content, and have "low reader reliability" (Mehrens & Lehmann, 1984, p. 98). "Low validity is also a problem with essay items. It has been said that not only do essay

items often not measure what they purport to, but that they take longer not to do it" (Bott, 1996, p. 136).

The following are some guidelines presented by Thorndike et al. (1991) in writing essay type items. First, have clearly in mind what mental process you want the student to use in responding to the question before starting to write it. Second, use novel material or organization of material in phrasing essay questions. Third, start essay questions with such words or phrases as "compare," "contrast," "give the reasons for," "give original examples of," "explain how," "predict what would happen if " "criticize," differentiate," and "illustrate." Avoid beginning essay questions with such words as "what," "who," "when," and "list" because these words direct students merely to reproduce information. Fourth, write the essay question in such a way that the task is clearly and unambiguously defined for each examinee. Fifth, a question dealing with a controversial issue should ask for and be evaluated in terms of the presentation of evidence for a position, rather than the position taken. Sixth, adapt the length and complexity of the answer to the maturity level of the students. Seventh, require all students to answer the same question. Eighth, provide an indication of how questions are weighted by indicating the number of points or the amount of time to be spent on each question.

The following are some poorly written essay type questions and their corresponding corrections.

1. Poor: What are Newton's laws of motion?

   Better: Describe each of Newton's three laws of motion. Illustrate each with the action of the ball in a game of baseball. (Nunnally, 1972, p.183)

2. Poor: What were the forces that led to the outbreak of the Civil War?

Better: Compare and contrast the positions of the North and South at the outbreak of the Civil War. Include in your discussion economic conditions, foreign policies, political sentiments, and social conditions. (Kubiszyn & Borich, 1984, p. 96)

Pool of items

In the building of a pool of items, the teacher assumes that there are certain objectives that will be stable and therefore can be listed, say on index cards, before any specific test is actually written. "The item is typed (or cut and pasted) on one side of the card with its objective, level of complexity from Bloom's taxonomy, grade and general subject content, and the source of the item" (Sax, 1974, p. 242). The correct answer should be indicated on the index card in some way, such as using an asterisk next to it or writing it on the back. From such index cards, it is possible to prepare a variety of test items that might be used in measuring the desired objectives. When it comes time to write a specific test, the teacher only needs to search the index cards for the items to include on the test. If the card file is complete, selecting the best items for the test will be done with a minimum of effort.

Conclusion

Test scores, when obtained from reliable and valid test items, can provide important information for both the teacher and the student. To the teacher, the test results will show how well the students mastered the material that was covered. Thus, the teacher can evaluate the effectiveness of teaching strategies. In addition, an item analysis might suggest to the teacher, which items not to use on subsequent administrations of the test. To the students, the test results will show them how much they know as well as how much they do not know. From this, the

students might decide to change their study habits. Therefore, since so many critical decisions are made based on the results of testing, the test constructor needs to be very careful when constructing a test.

References

Bott, P. A. (1996). Testing and assessment in occupational and technical education. Boston: Allyn & Bacon.

Ebel, R. & Frisbie, D. A. (1986). Essentials of educational measurement (4th ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Hopkins, K.D., & Stanley, J.C. (1981). Educational and psychological measurement and evaluation (6th ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Kubiszyn, T., & Borich, G. (1984). Educational testing and measurement. Glenview, IL: Scott, Foresman & Company.

Mehrens, W.A., & Lehmannn, I.J. (1984). Measurement and evaluation in education and psychology (3rd ed.). New York: Holt, Rinehart & Winston.

Nunnally, J.C. (1972). Educational measurement and evaluation (2nd ed.). New York: McGraw-Hill Book Company.

Oosterhof, A. (1994). Classroom applications of educational measurement (2nd ed.). New York: Macmillan Publishing Company.

Sax, G. (1974). Principles of educational measurement and evaluation. Belton, CA: Wadsworth Publishing Company, Inc.

Thorndike, R.M., Thorndike, R.L., Cunningham, G.K., & Hagen, E.P. (1991). Measurement and evaluation in psychology and education (5th ed.). New York: Macmillan Publishing Company.

Wiersma, W., & Jurs, S.G. (1990). Educational measurement and testing (2nd ed.). Boston: Allyn & Bacon.

Table 1

Solving Quadratic Equations

Performance Objective	Number of Questions
Factor Quadratic Equations	5
Solve Quadratic Equations	10
Apply Quadratic Equations	10

Table 2

Analyze and/or Solve Problems Involving Exponents, Quadratic Situations, or Right Triangles

	Taxonomy Level			
			Application	
Performance Objective	Knowledge	Comprehension	Synthesis, etc.	Total
Analyze and/or solve problems involving exponents the laws of exponents.	2	1	2 (20%)	5
Analyze and/or solve problems involving quadratic situations.	4	4	2 (40%)	10
Analyze and/or solve problems involving right triangles.	2	3	5 (40%)	10
Totals	8 (32%)	8 (32%)	9 (36%)	25